CHAPTER 14

Specifying Metadata

In This Chapter

META: The element

Name attributes

The http-equiv attribute

Robots.txt

This chapter is about how search engines track and index your page. The `META` tags you include have a big influence on how likely your page will be found by people who are using search engines to find specific answers. There are some building blocks to cover, as well as a few tips and tricks that may help.

The META Element

Meta means *about.* This element is the place where you give information about your page. Search engines use the `META` element to help index your page. The problem with the current scheme of `META` elements is that no standard exists. This is about to change and a standard is being developed. Using `META` elements according to the new standard is highly recommended.

Meta `<META>`

Start Tag: Required

Content: Empty

End Tag: Forbidden

Attributes: `http-equiv`: http response header name

`name`: metainformation name

`content`: value associated with name

`scheme`: select form of content usually used to provide helpful but noncritical information or to provide more context for the correct interpretation of metadata.

`lang`: add the desired language for the browser to read.

`dir`: select a direction for text to be read. For example, Arabic reads right to left.

Despite the fact that the META element is infinitely flexible, the way you use it is pretty straightforward. Here are some examples:

```
<META name="creator" content="Bill Karow">
<META name="publisher" content="IDG Books Worldwide">
<META name="description" content="really fabulous resource
about developing Web sites using all the cool new features of
HTML 4.0">
```

For the most part, when you use the META element, you assign a value to the name attribute and to the content attribute.

Name attributes

The Dublin Core Data Set (in Table 14-1) is the most widely accepted of the emerging standards, and you will be safe to draw the values of your name attributes from it. You certainly don't need to include all of these; they are optional. Also, if you need to include one more than once — say, there are multiple contributors — then you can repeat an element.

Table 14-1
The Dublin Core Data Set

Purpose	*Value of name attribute*	*Explanation*
Title	title	The title of the page.
Author or Creator	creator	The person or organization responsible for creating the page.
Subject and Keywords	subject	The subject of the page or keywords that describe the content of the page. As much as possible, you are encouraged to use controlled vocabularies and formal classification schemas (like the kind used by a library).
Description	description	A narrative description or abstract of the page.
Publisher	publisher	The publisher (your employer, your university, your publisher).
Other Contributor	contributor	Anyone who has contributed to the content of the page.
Date	date	The date of publication. Consider using YYYY-MM-DD. This is no place to get creative.

Purpose	*Value of name attribute*	*Explanation*
Resource Type	type	The type of page this is: novel, poem, personal page, and so on.
Format	format	Don't use this yet. A standard is under development.
Resource Identifier	identifier	A unique number you associate with the page so that no two documents at your URL have the same number.
Source	source	If your page was derived from a book, give the ISBN of the book. Otherwise, don't worry about this one.
Language	language	Language of the page. Use the 2-digit code, which can be found in Appendix F at the end of this book.
Relation	relation	This is an experimental attribute. If your page has a formal relationship with another page, this is where you can specify it.
Coverage	coverage	Perhaps to be used for geographical specification; they aren't sure yet.
Rights Management	rights	You can use this to link to a copyright notice or to specify your copyright statement.

Definition

ISBN. The International Standard Book Number is the unique 10-digit number assigned to a book by its publisher. For example, the ISBN of this book is 0-7645-3473-4. The number can usually be found on the back cover of a book or on the copyright page.

The http-equiv attribute

The previous name attributes are currently only good for search engines and people who want to view your source to see who else contributed to the creation of your fabulous page. The `http-equiv` attribute is something you can use instead of the name attribute that actually means something special to the browser that requested the page. The browser can use this data when displaying your page. A few examples are telling the browser when your content expires, the character set used in your page, and when to refresh the page:

```
<META http-equiv="expires" content="Mon, 23 Mar 1998 20:00:00
EST">
```

```
<META http-equiv="Content-Type" content="text/html;
charset=ISO-8859-5">
<META http-equiv="refresh"
content="10,http://www.mycompany.com/home.html">
```

In the last example, the refresh rate is in seconds. This is something you might use if you want to show a splash screen briefly and then have the browser automatically load the Welcome page.

Robots.txt

When a search-engine robot attempts to index a site, it looks for a file named robots.txt. The robots.txt standard is a very useful tool for both Webmasters and the people who run Web crawlers.

If the site `http://www.cyclingforum.com` is being indexed, the search engine looks for `http://www.cyclingforum.com/robots.txt`. This filename is case-sensitive, and must be lowercase. If the search-engine robot finds this document, it looks at it for instructions on what documents from that site, if any, it has permission to index. The robots.txt file can be written to disallow only certain search engines, or to protect specific directories from being indexed.

The following is a sample robots.txt file. The items beginning with a number sign (#) are for human viewing. You can say anything you like here, but I recommend something like the following:

```
# robots.txt file for http://yourdomain.com/
# e-mail webmaster@yourdomain.com for problems
# webmasters may change or alter this txt to fit your needs
User-agent: *
Disallow: /cgi-bin
Disallow: /images
Disallow: /logs
Disallow: contactus.html
```

The asterisk (*) indicates that the text to follow is for all search engines. See the list at the end of this chapter for specific robot names. The `Disallow:` command tells the robot not to index what follows. You can disallow entire folders or exact pages. Since Excite only accepts 25 pages per domain, you may want to use the `User-Agent` tag to make pages for specific search engines. `User-agent` is the name of the spider program.

In the robots.txt file, blank lines are permitted only between records. One `User-agent` field per record is allowed, though there can be as many user agent entries as needed, provided they are contained in separate records.

Save this text file in your top-level directory. It is the first page many search engines will seek and it is accepted by the major indexes. If you are using free space in someone else's directory (if you use AOL, for example) this will not work. Since there are so many pages inside the big services, most search engines ignore those sites. You should e-mail the various search engines asking to be indexed.

Some of the more common User-agents are:

- AltaVista — Scooter
- Excite — ArchitextSpider
- Infoseek — Infoseek Sidewinder
- Lycos — Lycos_Spider_(T-Rex)

From Here

Jump to Chapter 25 and get acquainted with style sheets.

Proceed to Chapter 15 and learn about structuring text with block-level elements and lists.

Summary

This chapter gave you a thorough explanation of search engines and how they look at pages. You learned about the somewhat confusing `META` element and all its variations. Finally, you learned how to write your HTML so that you can control if and how it will be indexed by most Web search engines and robots.

✦ ✦ ✦